# Thomas Bennett - trb090020

# CS 4395.001 - UTD Spring 2023

# Text Classification 1

***Poe's Law*, paraphrased: Any sufficiently advanced satire is indistinguishable from ~nuttery~ fake news.**

Golbeck et al. propose a dataset for training models to detect if a suspicious news article is fake news or satire.

## References

### Data

Author's Github Repo

### Paper

Fake News vs Satire: A Dataset and Analysis

In [1]:
```
import sklearn
import pandas as pd

df = pd.read_csv('./FakeNewsData/Fake-News-Stories.csv')
df.head()
```

Out[1]:

| | Article Number | URL of article | Fake or Satire? | URL of rebutting article | Fake or Satire?.1 |
|---|---|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | Fake | http://www.snopes.com/cdc-forced-vaccinations/ | Fake |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | Fake | http://www.snopes.com/white-house-logo-change/ | Fake |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | Fake | http://www.snopes.com/obama-veterans-money-to-... | Fake |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Fake | http://www.snopes.com/obama-veterans-money-to-... | Fake |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | Fake | http://www.snopes.com/california-to-jail-clima... | Fake |

## Get the text data

In [2]:
```
def get_data(article_num, fake):
    """Returns the text of the article"""

    location = './FakeNewsData/StoryText/'

    if int(fake) == 1:
        location += f'Fake/finalFake/{int(article_num)}.txt'
    else:
        location += f'Satire/finalSatire/{int(article_num)}.txt'

    with open(location, 'rb') as f:
        text = f.read().decode(errors='replace')
    return text

def get_fake(t):
    if t == 'Fake':
        return 1
    return 0
```

In [3]:
```
df.columns.values.tolist()
```

Out[3]:
```
['Article Number',
 'URL of article',
 'Fake or Satire?',
 'URL of rebutting article',
 'Fake or Satire?.1']
```

In [4]:
```
df.rename(columns={'Article Number': 'Number', 'URL of article': 'URL', 'Fake or Satir
df.drop(columns={'URL of rebutting article', 'Fake or Satire?.1'}, inplace=True)
df.dropna(inplace=True)
df.head()
```

Out[4]:

| | Number | URL | Target |
|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | Fake |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | Fake |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | Fake |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Fake |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | Fake |

### Note: Some rows had to be removed manually, as they referred to articles which were removed from the final dataset

In [5]:
```
import re

corpus = []
for i in range(len(df)):
    number = df.iloc[i][0]
    fake = get_fake(df.iloc[i][2])
    text = get_data(number, fake)
    text = " ".join(text.splitlines())
    corpus.append(text)
print(corpus[:2])
```

['CDC Proposes Rule to Apprehend and Detain anyone, anywhere, at any time, for any du
ration, without Due Process or right of Appeal - and administer FORCED Vaccinations!
http://www.redflagnews.com/headlines-2016/cdc-proposes-rule-to-apprehend-and-detain-a
nyone-anywhere-at-any-time-for-any-duration-without-due-process-or-right-of-appeal-an
d-administer-forced-vaccinations The Centers for Disease Control (CDC) has proposed a
"rule" giving them the power to apprehend and detain anyone, anywhere, at any time, w
ithout Due Process or any right of appeal, and to hold that person in quarantine for
as long as the CDC wants -- and no one can refuse them!  Editor\'s Opinion:  This is
the kind of tyranny that the Second Amendment is designed to protect Americans from.
Based on CDC\'s 8/15/16 publication of  \' Rules for the Control of Communicable Dise
ases\', the CDC is giving itself the power to forcibly apprehend healthy people  en m
asse, and detain them indefinitely with no process of appeal.  Kindly enough the CDC
is giving the public until 10/14/2016 to comment on its new found extra-Constitutiona
l power, "and whether there are any public concerns with the absence of a specific ma
ximum apprehension period in the regulation."  Of course and as would be expected fr
om a totalitarian unconstitutional power grab, "When an apprehension occurs, the indi
vidual is not free to leave or discontinue his/her discussion with an HHS/CDC public
health or quarantine officer."  Moreover, the CDC also would like the public\'s inpu
t on the fact their power is not limited to just individual persons but rather they c
ould apprehend entire cities in mass if they so desired: "HHS/CDC specifically reques
ts public comment on this proposed provision to issue Federal orders to entire groups
rather than individuals." And as is to be expected since its impossible to give a med
ical examine to an entire city, the CDC would also like your comments on the fact "th
e proposed practice to issue Federal orders before a medical examination has taken pl
ace. " For those wishing to give the CDC their requested comments on their new found
powers, the link to make such comments can be found under the SOURCE links at the end
of this article. Anyone who is interested would do well to read the CDC\'s entire pub
lication in the Federal Register.  The CDC\'s claimed power follows these Stages:  Y
ou (or your city) are declared "precommunicable"  Apprehension and Detention [A&D] Or
der of Isolation, Quarantine, or Conditional Release  In stage 1: "CDC defines precom
municable stage to mean the stage beginning upon an individual\'s earliest opportunit
y for exposure to an infectious agent", as previously indicated CDC does NOT need to
give a medical exam to declare you (or your city) "precommunicable". In fact, you may
be perfectly healthy, unexposed, and uninfected. All that is required is for someone
to say they suspect you (or your city) had a nebulous general and poorly defined "opp
ortunity" for exposure.  Moreover, you don\'t even get any due process to prove you h
ad zero opportunity to be exposed until after CDC has proceeded on to Stage 3; The ru
b being the CDC can hold you (or your city) at Stage 2 indefinitely with no appeal by
never proceeding to Stage 3. ie Do Not Pass Go, Go Directly To Jail.  In Stage 2: CD
C sneaks in its unlimited and unchecked authority. They\'ve couched this authority by
describing how they "generally" expect it to work in a temporary manner, but they\'ve
also clearly stated it is open ended and there is no discussion of due process in the
A&D phase (again CDC wants your comments on this fact).  To see how CDC\'s concept of
unlimited city wide Apprehension and Detention would play out, watch CDC\'s very own
2011 propaganda movieContagion. A movie basically written for and by the CDC to scare
the public into funding them and showcase how their heroic dream response to an outbr
eak would unfold.  There is no onus on the CDC to end the Apprehension phase, or to p
roceed on to issuing orders of isolation, quarantine, or conditional release; As such
you (or your city) can be held in apprehension and detention indefinitely.  In Stage
3: IF an order of isolation, quarantine, or conditional release is issued, the CDC gi
ves those so ordered one chance in the first 72 hours to ask CDC to change their mind
s, after which unlimited detention is again on the table.  The CDC can stop, detain,
and jail you anywhere.  The other key factors of note is that the CDC does NOT limit
this power to the international borders. CDC says it can take such actions any where
in the USA based on a claim that every action affects interstate travel.  CDC claims
it can set up check points at any bus or train station in the country, or at any loca
tion that might affect interstate travel.  CDC claims that the simple act of lining u
p at any CDC checkpoint gives them irrevocable authority to force you to be screened.
"an individual\'s refusal to be screened may result in quarantine, isolation, or cond
itional release" This could be in your car stuck in traffic at a CDC check point, a b

```
us station, a taxi-stand, etc etc.    "(holding that a passenger consents to an airpo
rt security search by presenting himself/herself for boarding and that such consent m
ay not be revoked by simply walking away). Thus, in order to protect interstate trave
l from communicable disease threats, HHS/CDC intends for this section to apply broadl
y to all circumstances where individuals may queue with other travelers"   "HHS/CDC b
elieves that the rationale for airport security screenings may be extended to other f
orms of transportation, e.g., trains and buses, because of the similar administrati
ve or special governmental need in preventing interstate communicable disease sprea
d"  FORCED VACCINATION   "CDC may enter into an agreement with an individual, upon su
ch terms as the CDC considers to be reasonably necessary, indicating that the individ
ual consents to any of the public health measures authorized under this part, includi
ng quarantine, isolation, conditional release, medical examination, hospitalization,
vaccination, and treatment; provided that the individual\'s consent shall not be cons
idered as a prerequisite to any exercise of any authority under this part." Even thou
gh the CDC believes it can force all the above procedures on individuals and groups,
it also uses "voluntary agreements" for the sole purpose of making forced actions eas
ier to perform. Anyone breaking these "voluntary agreements", even if they didn\'t ag
ree to them, is subject to criminal prosecution.  "individuals who violate the terms
of the agreement or the terms of the Federal order for quarantine, isolation, or cond
itional release (even if no agreement is in place between the individual and the gove
rnment), he or she may be subject to criminal penalties" Surprisingly, the one thing
CDC\'s left out of this rule making is the creation of the their own armed Federal Po
lice Force to carry out these action; but it can\'t be far off. This is simple:  If a
ny agency of any government tries to stop innocent, healthy people, without cause, wi
thout warrant and without charge, then tries to detain, quarantine or imprison you,
then what they are doing is unlawful denial of Constitutional rights and you have jus
t grounds to SHOOT THEM.  If you choose to shoot, make certain you SHOOT FIRST and SH
OOT TO KILL.', 'OUTRAGE: What Obama Just Did to the White House Logo Will Make You Si
ck http://www.redflagnews.com/headlines-2016/-outrage-what-obama-just-did-to-the-whit
e-house-logo-will-make-you-sick CONSERVATIVE TRIBUNE There are many conspiracy theori
es floating out there about President Barack Obama. Some claim he is a Muslim, a terr
orist or a Russian spy. However, what is known beyond any reasonable doubt is that he
is largely incompetent and has little love for America. Proof of this can be found by
looking no further that the White House logo, which was redesigned in 2016. Everyone
known what the White House logo is supposed to look like  a portrait of the North Fa
ce of the White House  but what few people have noticed over the past few years was
the change at the top of the logo. For many years, the White House logo had the Ameri
can flag flying on the top of the White House  as it does in real life. In Obama s
redesigned version, there is no American flag, just a white flag. Here is what the or
iginal logo looked like:  And here is Obama s sanitized version:  The white flag is a
common symbol for surrender, which has many people wondering if Obama was trying to s
ecretly signal to America s enemies that he was surrendering. In all seriousness, thi
s probably wasn t some secret signal. It was just Obama disrespecting America and tr
ying to get rid of anything that made us special  which really shouldn t surprise a
nyone at this point. Obama has disgraced the American flag multiple times while in of
fice, so it should be no surprise that he would completely remove it from the White H
ouse logo. We ve suffered for almost eight years under a president who has made it cl
ear he doesn t think America is special. We desperately need to elect someone who wil
l restore our faith in our country and put America, and Americans, first. H/T WZ']
```

In [6]:
```
len(df) == len(corpus)
```

Out[6]:
```
True
```

In [7]:
```
df['Text'] = corpus
df.head()
```

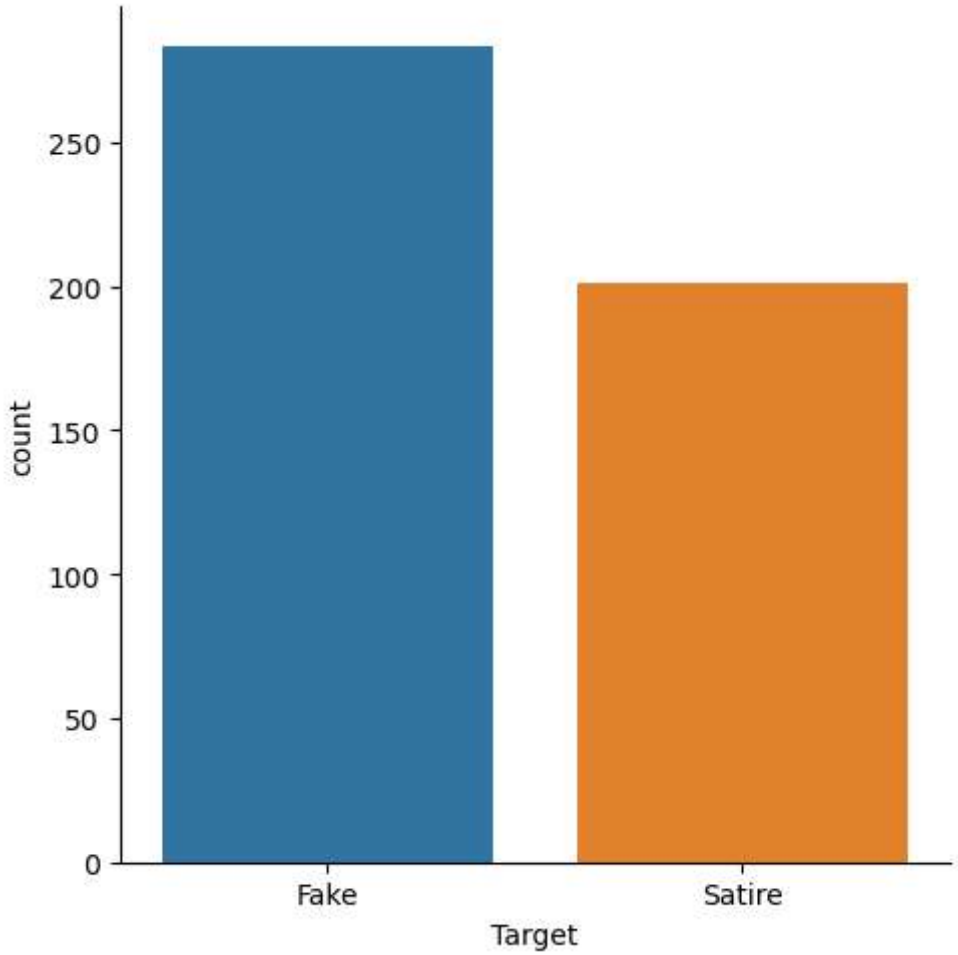
Out[7]:

| | Number | URL | Target | Text |
|---|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | Fake | CDC Proposes Rule to Apprehend and Detain anyo... |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | Fake | OUTRAGE: What Obama Just Did to the White Hous... |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | Fake | White House CANCELS all Obama Appearances at H... |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Fake | Obama Cuts 2.6 Billion From Veterans While All... |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | Fake | California Introduces Law To Jail Anyone Who Q... |

In [8]:
```
import seaborn as sb

sb.catplot(x='Target', kind='count', data=df)
```

Out[8]:
```
<seaborn.axisgrid.FacetGrid at 0x26671519690>
```




Since the two classes are not terribly unbalanced, I will not bother applying a tradeoff.
Satire=201, Fake=283.
I will set up the model to predict whether an article is Fake, so an accuracy of better than 58% (283/484) will be an improvement over sheer luck.

In [9]:
```
dummy = pd.get_dummies(df['Target'])
df = pd.concat([df, dummy], axis='columns')
df.head()
```

Out[9]:

| | Number | URL | Target | Text | Fake | Satire |
|---|---|---|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | Fake | CDC Proposes Rule to Apprehend and Detain anyo... | 1 | 0 |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | Fake | OUTRAGE: What Obama Just Did to the White Hous... | 1 | 0 |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | Fake | White House CANCELS all Obama Appearances at H... | 1 | 0 |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Fake | Obama Cuts 2.6 Billion From Veterans While All... | 1 | 0 |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | Fake | California Introduces Law To Jail Anyone Who Q... | 1 | 0 |

In [10]:
```
df.drop(columns={'Target', 'Satire'}, inplace=True)
df.head()
```

Out[10]:

| | Number | URL | Text | Fake |
|---|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | CDC Proposes Rule to Apprehend and Detain anyo... | 1 |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | OUTRAGE: What Obama Just Did to the White Hous... | 1 |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | White House CANCELS all Obama Appearances at H... | 1 |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Obama Cuts 2.6 Billion From Veterans While All... | 1 |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | California Introduces Law To Jail Anyone Who Q... | 1 |

In [11]:
```
df.rename(columns={'Fake': 'Target'}, inplace=True)
df.head()
```

Out[11]:

| | Number | URL | Text | Target |
|---|---|---|---|---|
| **0** | 375 | http://www.redflagnews.com/headlines-2016/cdc-... | CDC Proposes Rule to Apprehend and Detain anyo... | 1 |
| **1** | 376 | http://www.redflagnews.com/headlines-2016/-out... | OUTRAGE: What Obama Just Did to the White Hous... | 1 |
| **2** | 377 | http://www.redflagnews.com/headlines-2016/whit... | White House CANCELS all Obama Appearances at H... | 1 |
| **3** | 378 | http://www.redflagnews.com/headlines-2016/obam... | Obama Cuts 2.6 Billion From Veterans While All... | 1 |
| **4** | 379 | http://www.redflagnews.com/headlines-2016/cali... | California Introduces Law To Jail Anyone Who Q... | 1 |

## Naive Bayes

In [12]:
```
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split

stopwords = stopwords.words('english')

X = df['Text']
y = df['Target']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15)
vectorizer = TfidfVectorizer(stop_words=stopwords)

X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)

print(f'Train size: {X_train.shape}')
print(f'Test size: {X_test.shape}')
```

```
Train size: (411, 15155)
Test size: (73, 15155)
```

In [13]:
```
from sklearn.naive_bayes import MultinomialNB

naive_bayes = MultinomialNB()
naive_bayes.fit(X_train, y_train)
```

Out[13]:
```
MultinomialNB()
```

In [14]:
```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, c

preds = naive_bayes.predict(X_test)

print(confusion_matrix(y_test, preds))
print(f'Accuracy: {accuracy_score(y_test, preds)}')
```

```
[[ 2 27]
 [ 0 44]]
Accuracy: 0.6301369863013698
```

After doing several runs the model is unable to do better than natural chance (about 60%) and often does much worse (as low as 40%). Clearly the model has developed an extremely strong tendency to simply guess that everything is satire, becuase it calculates very few features to be fake news.

## Logistic Regression

In [16]:
```
from sklearn.linear_model import LogisticRegression

classifier = LogisticRegression(C=10.0, solver='newton-cholesky')
classifier.fit(X_train, y_train)
preds2 = classifier.predict(X_test)

print(confusion_matrix(y_test, preds2))
print(f'Accuracy: {accuracy_score(y_test, preds2)}')
```

```
[[19 10]
 [ 5 39]]
Accuracy: 0.7945205479452054
```

I set C to a high value because I believe the Naive Bayes model was overfitted. The scikit-learn documentation for LogisticRegression recommends using the Newton-Cholesky solver for feature matrices with num_samples > num_features.

## Neural Network

In [18]:
```
from sklearn.neural_network import MLPClassifier

classifier = MLPClassifier(solver='lbfgs')
classifier.fit(X_train, y_train)
preds3 = classifier.predict(X_test)

print(confusion_matrix(y_test, preds3))
print(f'Accuracy: {accuracy_score(y_test, preds3)}')
```

```
[[16 13]
 [ 5 39]]
Accuracy: 0.7534246575342466
```